Identifying Proteins Involved in Parasitism by Discovering Degenerated Motifs

Celine Vens[1,2], Etienne Danchin[2], and Marie-Noëlle Rosso[2]

[1]Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, 3001 Leuven, Belgium
celine.vens@cs.kuleuven.be
[2]Institut National de la Recherche Agronomique
400 route des Chappes, BP 167, 06903 Sophia-Antipolis Cedex, France
{etienne.danchin,rosso}@sophia.inra.fr

1 Introduction

Identifying motifs in biological sequences is an important challenge in biology. Proteins involved in the same biological system or physiological function (e.g., immune response, chemo-sensation, secretion, signal transduction,...) are subject to similar evolutionary and functional pressures that have an outcome at the protein sequence level. Finding motifs specific to proteins involved in the same process can help deciphering the determinants of their fate and thus be used in identifying new candidate proteins involved in important biological systems.

To our knowledge all currently available methods search motifs in protein sequences at the amino acid level, sometimes allowing degenerate motifs to comply with point variations [1, 2]. However, it is known that conservation of the three-dimensional structure is more important than conservation of the actual sequence for the biological function and proteins that have no detectable sequence similarity can fold in similar structures. At a given position in the sequence, the nature and physico-chemical properties of amino acids in protein families is more conserved than the amino acid itself.

We propose a method that allows to identify emerging motifs based both on conservation of amino acids and on the physico-chemical properties of these residues. Given a set of protein sequences known to be involved in a common biological system (positive set) and a set of protein sequences known not to be involved in that system (negative set) our method is able to identifiy motifs that are frequent in positive sequences while infrequent or absent in negative sequences. The identified motifs can then be used to mine the wealth of protein data now available, in order to identify new previously uncharacterized proteins involved in biological processes of importance.

In this work, the biological system of interest is the protein secretion of a plant parasitic nematode (roundworm). The nematode in question, *Meloidogyne incognita* [3], is a major crop devastator, and controlling it has become an important issue. In this context, it is important to identify the proteins secreted by the nematode into the plant (e.g. cell-wall degrading enzymes that allow the parasite to enter the plant).

2 Identifying Degenerated Amino Acid Patterns

2.1 Formal Task Description

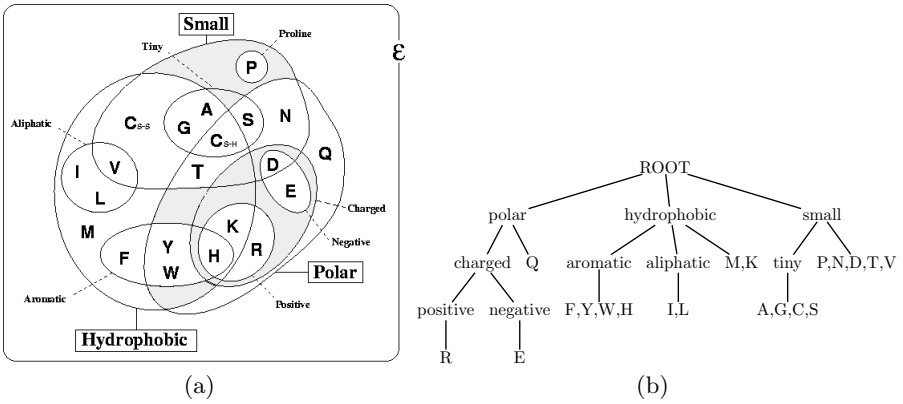
We define the task of identifying degenerated emerging protein motifs as follows:

Given: (1) a set of positive proteins P, and a set of negative proteins N, (2) two frequency thresholds F_P and F_N, (3) a set of physico-chemical amino acid properties C and a partial order $\preceq$ defined on the union of C and the amino acid alphabet A. For all $ca_1, ca_2 \in C \cup A$: $ca_1 \preceq ca_2$ if and only if ca_1 is more general than ca_2.

Find: the set of all patterns M, using symbols in $C \cup A$, that have $freq(M, P) \geq F_P$ and $freq(M, N) \leq F_N$. The function $freq(X, Y)$ returns the number of proteins in set Y that contain the pattern X.

2.2 Classification scheme

Several amino acid classifications exist in the literature. In this work, we use the classification by Russell et al [4]. It describes amino acids according to their hydrophobicity, size, and polarity, see Fig. 1(a).



(a) (b)

Fig. 1. (a) Venn diagram of amino acid properties [4]. (b) Spanning tree.

2.3 Algorithm

The algorithm we propose is based on the well-known generate-and-test principle, introduced in the Apriori algorithm [5]. At each iteration, a set of candidates is generated, whose frequency is tested. Given the partial order mentioned above,

the search space of all possible patterns is structured as a lattice, with an artificial root element that denotes the empty pattern. The lattice represents an ordering relation: a pattern $(p_1, p_2, p_3, ..., p_n)$ is more general than another pattern $(q_1, q_2, q_3, ..., q_m)$, if and only if $n \leq m$ and for each pair (p_i, q_i) it holds that $p_i \preceq q_i$.

In order to conduct the search efficiently, the candidate generation exploits the antimonotonicity properties of the frequency constraints. This results in the following rules:

- If for a pattern M it holds that $freq(M, P) \leq F_P$, then the pattern does not need to be specialized, since for all its children C, it will hold that $freq(C, P) \leq F_P$.
- If for a pattern M it holds that $freq(M, N) \leq F_N$, then for all its children C, it will hold that $freq(C, N) \leq F_N$, we do not need to test them.

We have implemented the algorithm using a depth-first search strategy. Essentially, the algorithm looks for those patterns that are frequent in the positive sequences, and meanwhile checks if they are infrequent in the negative sequences. We discuss its most important parts.

Candidate generation. In order to perform a complete search, it is important that each relevant pattern in the lattice is considered, and that no pattern is considered more than once. To achieve this, our candidate generation method traverses the lattice from general to specific, and at each step performs two basic operations to generate new candidates given a pattern:

- add a top-level element of the partial order
- minimally specialize the last element of the pattern

In order to ensure that no pattern is considered more than once, we first construct a spanning tree out of the partial order DAG (see Fig 1(b)), and specialize the pattern using this tree.

Candidate pruning and testing. When testing a candidate, it is not necessary to check the complete set of positive sequences, it suffices to check the sequences containing the parent candidate, and only in the case all parents have passed the minimal frequency threshold F_P. In order to exploit this property, we have to make sure that all parents have been tested before a pattern is considered, i.e. the spanning tree of the amino acids and their properties has to be constructed in a way that, in depth-first traversal, all parents of a node are visited before the node itself is visited. The tree shown in Fig. 1(b) fulfils this constraint.

3 Results

We have generated a set of 100 *M. incognita* proteins, that were experimentally proven to be secreted into plants. As negative set, we took 130 proteins that

are conserved in non-parasitic nematodes, i.e. that are unlikely to be involved in parasitism.

As we are interested in identifying motifs that are specific to secreted proteins, the maximal frequency threshold for the negative set, F_N, was set to 0, i.e., we look for so-called jumping patterns. The minimal frequency threshold for the positives, F_P, was set to 20.

The algorithm has identified 3 motifs:

- (hydrophobic charged polar small hydrophobic small hydrophobic tiny small small hydrophobic)
- (hydrophobic hydrophobic small polar polar hydrophobic T hydrophobic polar small hydrophobic hydrophobic)
- (small small small small polar small tiny hydrophobic polar polar hydrophobic small)

Together, these motifs cover 40 of the secreted proteins. Six proteins, including 2 plant cell-wall degrading (PCWD) enzymes, contain all 3 motifs. If we search the complete proteome of *M. incognita*, consisting of 19212 proteins [3], for proteins that contain the 3 motifs (assuming that these would be the most probable of being putative secreted proteins), we obtain a set of 43 proteins that were not included in the training set. Among these, we observe 4 extra PCWD enzymes, which have not yet been experimentally shown to be secreted. We are currently looking into the rest of the set.

4 Conclusions

We have proposed an algorithm for the identification of protein motifs that are not restricted to a sequence of amino acids, but can involve physico-chemical amino acid properties. The algorithm uses a traditional generate-and-test approach, with a specific candidate generation operator and pruning step. The algorithm was applied to the task of identifying motifs specific to the secreted proteins of a plant-parasitic nematode, resulting in three degenerate motifs, and a list of 43 candidate proteins to be tested for their involvedness in parasitism.

References

1. Bailey, T., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press (1994) 28–36
2. Ji, X., Bailey, J.: An efficient technique for mining approximately frequent substring patterns. In: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, IEEE Computer Society (2007) 325–330
3. Abad, P.e.a.: Genome sequence of the metazoan plant-parasitic nematode meloidogyne incognita. Nat Biotechnol. **26**(8) (2008) 909–915
4. Betts, M., Russell, R.: Amino acid properties and consequences of subsitutions. In: Bioinformatics for Geneticists. Wiley (2003)
5. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996)